# Lecture 21: Point Estimation

24 March 2016

So far we have discussed Mean Square Error performance of estimators. In this lecture we shall see the loss function framework for evaluation of estimators.

## 1 General Loss Function Framework Ingredients

1. Parameter space: $\Theta$ (e.g. $\mathbb{R}$)
2. Observation space: $\mathscr{X}$
3. Family of distributions indexed by $\Theta$: $\{f(x|\theta),\ \theta \in \Theta\}$
4. Action/Decision/Output space: $\mathscr{A}$
   (typically $\mathscr{A} \supseteq \Theta$, because estimator can give output $\notin \Theta$)
5. Loss function
   L: $\Theta \times \mathscr{A} \to \mathbb{R}_+$
   L($\theta$,a): "cost" suffered when estimating $\theta$ to be equal to a.
   (Ideally, if $\mathscr{A} = \Theta$; then L($\theta$,a)=0 when a=$\theta$)

   Given below are some examples of loss functions.
   Assume $\Theta = \mathscr{A} = \mathbb{R}$

   (a) Absolute loss
   $$L(\theta,a) = |\theta - a|$$
   (b) Square loss (corresponds to MSE)
   $$L(\theta, a) = (a - \theta)^2$$
   (c) Zero-One loss
   $$L(\theta, a) = \mathbb{1}_{\{\theta \neq a\}}$$
   (d) p-norm loss
   $$L(\theta, a) = |\theta - a|^p$$

Given an estimator W(X), (W: X$\rightarrow \mathscr{A}$) of $\theta \in \Theta$, {X$\sim$ f(x|$\theta$)}, its RISK FUNCTION at $\theta \in \Theta$ is given as :

$$R(\theta,W) = \mathbb{E}_\theta[L(\theta, W(X))]$$
$$= \int_{\mathscr{X}} L(\theta, W(X)).f(x|\theta)dx$$

(If L is square loss, then the above risk R gives the mean square error)

Our goal is to design W to minimize R($\theta, W$) over "all or most $\theta \in \Theta$".

Now given two estimators over the parameter space $\Theta$, how do we compare their performance and choose the best?

Consider the figure shown below. The x-axis represents the parameter space $\theta \in \Theta$ and y-axis represents the risk, $R(\theta, W)$ for an estimator W w.r.t $\theta$.

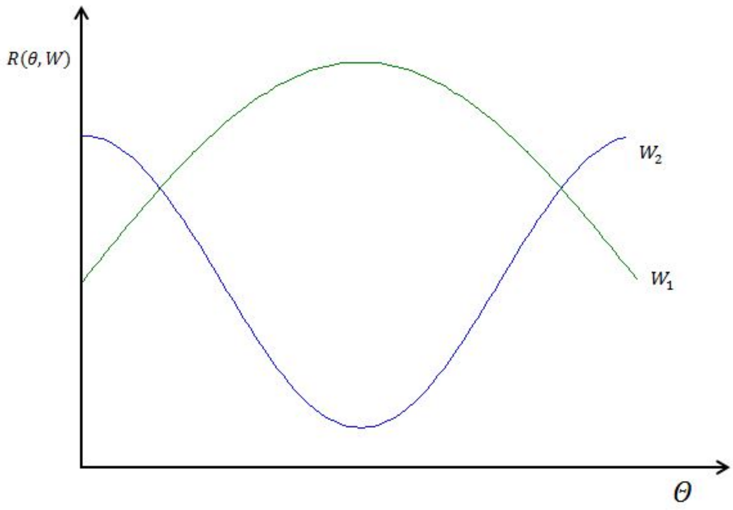


Figure 1: Risk v/s $\Theta$ for different estimators

One way to decide on the best estimator $W^*$ would be to choose the one having smaller peak. One can see that this is equivalent to the minimax estimator (as we are choosing the W with minimum $\max_\theta R(\theta, W)$ ).

Another option is to choose W that minimizes the area under the R(.,W) function. This is equivalent to the Bayesian estimator.

# 2 Notions of Optimality (Rule to compare estimators)

1. Bayes Risk :
   Asume a prior probability distribution $\pi$ over the parameter space $\Theta$ is given.
   The bayes risk of W = W(x) is
   $B_\pi(W) = \int_\Theta R(\theta, W).\pi(\theta)d\theta$
   Any estimator W that minimizes $B_\pi(.)$ over all estimators is called a Bayes estimator(denoted by ${W^*}_\pi$)

2. Max Risk(No prior necessary):
   $\overline{R}$(W) = $\sup_{\theta\in\Theta} R(\theta, W)$
   Estimator minimizing $\overline{R}(.)$ are minimax estimators.

## 2.1 Bayes Estimators

Bayes risk under prior $\pi$:

$$\begin{aligned}
B_\pi(W) &= \int_\Theta R(\theta, W)\pi(\theta)d\theta && (1)\\
&= \int_\Theta \int_{\mathscr{X}} L(\theta, W(X)).f(x|\theta)dx.\pi(\theta)d\theta && (2)\\
&= \int_{\mathscr{X}} \left[\int_\Theta L(\theta, W(X))\pi(\theta|x)d\theta\right] m(x)dx && (3)
\end{aligned}$$

where we have used $f(x|\theta).\pi(\theta) = \pi(\theta|x).m(x)$

and we have defined
m(x) $\equiv$ marginal of x
$= \int_\Theta \pi(\theta').f(x|\theta')d\theta'$

$\pi(\theta|x) \equiv$ Posterior density of $\theta$ given x
$= \dfrac{\pi(\theta).f(x|\theta)}{m(x)}$

Note that the quantity [.] is a function of only x ( and not $\theta$)

that implies, to minimize $B_\pi(W)$, we should choose
$\forall$x $\in \mathscr{X}$ : W(X) $\in argmin_{a\in\mathscr{A}} \int_\Theta L(\theta, a)\pi(\theta|x)d\theta$

i.e., a Bayes estimator minimizes the posterior expected loss given the data x.

**Example 2.1 (Bayes estimator for square-loss function).** Let $\Theta = \mathscr{A} = \mathbb{R}$
$L(\theta, a) = (a - \theta)^2$
The posterior expected loss is
$\int_{\mathbb{R}} (a - \theta)^2 \pi(\theta|x) dx$
Then the Bayes estimator is $W(X) = \int_{\Theta} \theta \pi(\theta|x) dx$
i.e., the posterior mean.

**Example 2.2 (Bayes estimator for absolute loss function).** Let $\Theta = \mathscr{A} = \mathbb{R}$
$L(\theta, a) = |a - \theta|$
The posterior expected loss is
$\int_{\mathbb{R}} |a - \theta| \pi(\theta|x) dx$
Here the Bayes estimator returns $W(X) = MEDIAN(\pi(.|x))$

*Proof.* The posterior expected loss is given by
$\mathbb{E}|x - a| = \int_{\mathbb{R}} |x - a| \pi(\theta|x) dx = \int_{-\infty}^{a} -(x - a)\pi(\theta|x) dx + \int_{a}^{\infty} (x - a)\pi(\theta|x) dx.$
The bayes estimator is given by
W(x) = $argmin_a \mathbb{E}|x - a|$.
Minimum can be obtained by computing the derivative and equating to 0.
$\frac{d}{da}\mathbb{E}|x - a| = \int_{-\infty}^{a} \pi(\theta|x) dx - \int_{a}^{\infty} \pi(\theta|x) dx$
Equating this equation to zero gives the result as $a = MEDIAN(\pi(.|x))$ □

(Similarly a 0-1 loss function returns $W(X) = MODE(\pi(.|x))$ )

## 2.2 Minimax Estimator

It turns out that minimax estimation is complicated. The main takeaway here is that the bayes estimator with constant risk over $\Theta$ is minimax.

**Definition 2.3.** A prior $\pi$ over $\Theta$ is a **LEAST FAVORABLE PRIOR** if it has the highest bayes risk,i.e
$B_{\pi}(W^*_{\pi}) \geq B_{\pi\prime}(W^*_{\pi\prime})$ $\forall$ prior $\pi\prime$ on $\Theta$ .

**Theorem 2.4.** *Suppose W is the Bayes estimator for some prior $\pi$ over $\Theta$, if $L(\theta, W)$ is a constant $\forall\ \theta \in \Theta$, then*

1. *$\pi$ is a least favorable prior*
2. *W is a minimax estimator.*

# 3 Asympotic Evaluation of Estimators

The goal here is to study what happens to the quality of estimation as the number of samples tend to infinity.

**Definition 3.1.** Let $W_n \equiv W_n(X_1, ..., X_n)$ for n $\geq 1$, be a sequence of estimators, for $\theta$, and assuming $X_i \overset{iid}{\sim}$f(x|$\theta$), then $W_n$ is **CONSISTENT** for estimating $\theta$ if $\forall \theta \in \Theta$, $W_n \overset{P_\theta}{\to} \theta$.
i.e $\forall\ \theta \in \Theta, \epsilon > 0$, $lim_{n \to \infty}$P[| $W_n$ - $\theta| \geq \epsilon] = 0$.

**NOTES**

1. Consistency is equivalent to convergence to quantity being estimated.

2. Need convergence in probability $\forall\ \theta \in \Theta$

Since mean-square convergence implies convergence in probability,
$\forall\ \theta \in \Theta$, $E_\theta[(W_n - \theta)^2] \to \infty$ as n $\to \infty$ is enough to show that $\{W_n\}$ is consistent.

**Theorem 3.2.** *If $W_n \equiv W_n(X_1, ..., X_n)$ is sequence of estimators such that $\forall\ \theta$,*

1. *$lim_{n \to \infty} var_\theta[W_n] = 0$*

2. *$lim_{n \to \infty} \mathbb{E}_\theta[W_n] - \theta = 0$*

*then $\{W_n\}$ is consistent.*

**Example 3.3 (Consistency of sample mean).** Let $X_1, ...., X_n \overset{iid}{\sim}$ f(x|$\theta$),for $\theta \in \Theta \subseteq \mathbb{R}$, and $\forall\ \theta \in \Theta$, $\mathbb{E}_\theta[|X_1|] < \infty$,let $W_n = \frac{1}{n}\sum_{i=1}^{n} X_i$; $\forall$ n $\geq 1$:
$\{W_n\}$ is consistent for estimating $\mathbb{E}_\theta[X]$ since, $\frac{1}{n}\sum_{i=1}^{n} X_i \overset{P_\theta}{\to} \mathbb{E}_\theta[X_1]$ = g($\theta$), due to the Weak Law of Large Numbers.

## 3.1 Consistency of Maximum Likelihood Estimator

**Recall** $X_1, ...., X_n \overset{iid}{\sim}$ f(x|$\theta$),for $\theta \in \Theta \subseteq \mathbb{R}$, MLE of $\theta$ is $argmax_{\theta \in \Theta} \prod_{i=1}^{n}$f($x_i|\theta$) or we can say,
$W_{MLE} \in argmax_{\theta \in \Theta} \sum_{i=1}^{n}$log(f($x_i|\theta$)).

**Theorem 3.4 (Consistency of MLE).** *Suppose $X_1, ...., X_n \overset{iid}{\sim}$ f(x|$\theta$), for $\theta \in \Theta \subseteq \mathbb{R}$,and f(x|$\theta \in \Theta$) satisfies some regularity conditions, then $\forall \theta \in \Theta$, $W_{MLE}^{(n)} \overset{P_\theta}{\to} \theta$.*